从 ChatGPT 到 AGI:生成式 AI 的媒介特质与伴生风险（八）

二、生成式 AI 的伴生风险

7. 人为操纵导致编造“真实性”，会制造虚假叙事操纵政治实施认知战。兰德高级工程师克里斯托弗·莫顿（Christopher Mouton）在 C4ISRNET 网站撰文，分析 ChatGPT 等大型语言模型带来的国家安全新风险，指出人工智能生成内容会展现出一种被称为“真实性”的现象。[1]“真实性”现象与自信“幻觉”不同，是人为刻意利用人工智能来编造虚假信息。尽管内容缺乏事实，但人工智能编造内容的逻辑性迷惑性较强，超出一般谣言的水准。利用人工智能编造听起来、看起来“真实”的信息，创作有说服力的内容，可以获得一定传播优势。2024 年 1 月 20-21 日，美国新罕布什尔州居民接收到一条“来自美国总统拜登”的政治性质电话语音。该条语音模仿拜登的声音，以拜登口头禅“真是一派胡言”开头，要求居民在初选期间留在家中，不要参与 1 月 23 日的初选投票，不要给特朗普投票，而是把选票留到 11 月大选时投给民主党。随后，白宫新闻秘书皮埃尔澄清说，这是一则伪造的电话录音。类似情况不断出现。1 月 21 日，一段模仿曼哈顿民主党领袖基思·赖特（Keith Wright）的音频流传，其中还混杂了模仿民主党议员伊内

兹·狄更斯（Inez Dickens）的声音。不法分子更倾向伪造音频，因为视频更易被识别造假。分析人士担心，在美国选民容易受到错误信息影响的当下，人工智能可能会在大选期间制造出更多混乱。据不完全统计，2024 年全球将有 70 多个国家或地区举行重要选举，覆盖超过全球半数人口。类似情况不断出现。2024 年，人工智能深度伪造生成欺骗性内容干扰选举被认为是全球面临的重要挑战。必须看到，人工智能生成的内容是多模态的，但审查方法和技术滞后，有害信息难以被发现。在第 60 届慕尼黑安全会议上，全球多家科技企业就签署协议，承诺将在 2024 年打击旨在干扰选举的人工智能滥用行为。

有学者研究发现，深度伪造应用在制造认知混淆和引发情绪反应方面具很强影响力，引发复杂的整治、经济、社会和文化等一系列问题。何康等人研究了深度伪造技术在现代战争宣传中的使用，发现深度伪造能够造成认知混淆，引发情绪反应，并在国际社交媒体上制造出“真与假”的认知效应。[2] Weikmann 等人也发现，人工智能生成的虚假视频在政治应用和社会影响方面的实际作用相对有限，而“去情境化”的图像和视频被视为更大的威胁。[3] 未来国家间的信息作战，也会利用利用人工智能为虚假信息披上可信的欺骗性外衣，进而影响对方的政治制度和社会秩序。例如，俄罗斯黑客伪造乌克兰总统泽连斯基呼吁乌军民放下武器的视

频。2023 年 3 月，微软研究院就发布《人工通用智能的火花：GPT-4 的早期实验》报告指出，像任何强大的技术一样，ChatGPT 可以被用来实施破坏行动，从有效生成虚假信息到发动网络攻击。ChatGPT 的交互能力可以被用来操纵、劝说或影响人们，包括构建虚假信息，生成用于说服的内容。例如，ChatGPT 能够一个创建虚假信息计划，包括确定分享虚假信息的网络平台、寻找与个人分享的信息来源以及确定使用情感诉求进行说服等策略。它还可以通过定制能够引发不同情绪反应的信息来实现思想意识干扰。再如，ChatGPT 可以利用社交媒体来传播故意侮辱、歧视个人或群体等信息。不良企图者能够利用 ChatGPT 生成虚假新闻或者谣言，通过邮件、社交媒体帖子等分发，通过富有说服力的文本影响人们的思想认知。尽管 ChatGPT 本身并不是实体武器，但与所有新技术一样，极有可能被武器化。可以预见，ChatGPT 会被一些国家或组织用来制造虚假或混乱内容，操纵公众舆论，侵略或攻击其他文化等[4]，沦为认知战工具。如今，已有不良企图者滥用 ChatGPT 大规模生成虚假信息，如伪造评论，编写有违是非曲直、指鹿为马的虚假信息。可见，人工智能能够生产大量虚假或误导信息，并被用于舆论操控等行为，如利用深度伪造合成假新闻、假数据以激化国际矛盾。

案例：大模型与认知战

OpenAI 虽未在面上被直接用于杀人或造成破坏，但其

应用所发动的认知战不容忽视。所谓“认知战”（Cognitive Warfare，CogWar）是指利用认知的各个方面来扰乱、破坏、影响或改变人类的决策。ChatGPT 基本具备类人化的学习能力和逻辑常识，从发帖模式到用户信息都与真实用户高度相似，可模拟人类的复杂思维并使用高质量且带有情感色彩进行互动，进而能通过投放“信息炸弹”混淆公众视听、塑造虚拟意见领袖。随着用户对大模型的依赖程度加深，大模型技术可以实现更隐性的认知控制。

·信息扭曲与认知操控：自动化生成虚假信息，编写误导性的文章或评论，操纵社交媒体内容，大模型通过大规模的信息搜集和分析，并针对个人需求和偏好，揭示其身份和行为，从而对个人进行针对性的攻击或操纵，或者操控其舆论和意识形态，导致信息过滤泡沫的形成，加剧社会的分裂和对立。俄乌冲突期间，一段关于乌总统泽连斯基宣布投降的伪造视频曾在社交平台疯传。

·煽动暴力与 AI 武器：通过操纵舆论和信息，大模型可以被用来煽动暴力行为，激化民族、宗教或政治之间的紧张关系，如鼓励恶意行为或组织暴乱。这可能导致人们受伤甚至丧生。如果大模型应用于军事领域，可能导致人工智能武器滥用和误用。这可能引发战争、冲突和人员伤亡，严重威胁生命安全。例如，乌安全局曾关闭一处拥有百万机器人的“巨魔工厂”，工厂目标就是抹黑乌官方发布的信息，破

坏乌社会与政治局势、伺机制造内部冲突。

[1] “真实性”一词是由电视台主持人斯蒂芬·科尔伯特在21世纪初创造出来的，用来描述信息是如何感觉正确的。这一概念强调，尽管缺乏事实的准确性，但具有高度连贯逻辑结构的内容可以影响聪明、老练的人决定事情的真伪。

[2] 何康,张洪忠,刘绍强等.认知的罗生门效应制造:深度伪造在俄乌冲突中的案例分析[J].新闻界,2023(01):88-96. DOI: 10. 15897/j. cnki. cn51-1046/g2. 20221208. 002.

[3] Weikmann T, Lecheler S. Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor-Network. Digital Journalism, 2023, 1-18.

[4] 王文广.跨文化传播中的通用人工智能：变革、机遇与挑战[J].对外传播,2023(05):48-51.